UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MASSACHUSETTS

SINGULAR COMPUTING LLC, Plaintiff, v. GOOGLE LLC, Defendant.	C.A. No. 1:21-cv-12110-FDS

DEFENDANT GOOGLE LLC'S MEMORANDUM OF LAW IN SUPPORT OF ITS RULE 12(B)(6) MOTION TO DISMISS FOR FAILURE TO STATE A CLAIM FOR PATENT INFRINGEMENT

TABLE OF CONTENTS

TABLE OF AUTHORITIES

CASES

RULES

I. INTRODUCTION

Singular has taken the "rinse, repeat" approach to patent litigation: it applied for new claims in the patent office well after the accused products were released and, after its original suit ran into challenges, sued again. Singular now asserts patent claims that were not drafted in 2009, when Singular filed its first patent application, or at any time in the ensuing decade. Instead, the applications that matured into the asserted patents were first filed in the PTO in 2020—years after Google's 2017 public debut of the accused TPU v2, and after Singular's original December 2019 suit was filed.

Singular's new claims, despite relying on the same specification, abandon what Singular previously told the Court was the touchstone of the invention Dr. Bates disclosed in that specification. Previously, in opposing dismissal on patentability grounds, Singular described the purported novelty of Dr. Bates' invention as "LPHDR computers with execution units (circuits) . . . generating *materially inaccurate results.*"[1] Singular again emphasized this now-abandoned point of novelty at the first suit's hearing on Google's motion to dismiss: "This specifically is designed and made in the chip to have a *certain specific amount of error,* not just general error. *The error is specific*, and the specificity of the error, the specificity of this is *to optimize the result.*"[2] The new claims Singular now asserts contain no limitation directed to the amount of error in outputs as in the previously asserted claims, thus abandoning altogether the purportedly key inventive feature of error and materially inaccurate results.

---

[1] Pl.'s Opp'n to Def.'s Rule 12(b)(6) Mot. to Dismiss for Lack of Patentable Subject Matter at 17, Case No. 19-cv-12551-FDS ("*Singular I*") (D. Mass. May 15, 2020), ECF No. 44. The Court may take judicial notice of proceedings on its own docket. *Perez-Tino v. Barr*, 937 F.3d 48, 54 n.3 (1st Cir. 2019) (citing *LeBlanc v. Salem (In re Mailman Steam Carpet Cleaning Corp.*), 196 F.3d 1, 8 (1st Cir. 1999)).

[2] Mot. to Dismiss Hr'g Tr. at 25:22-25, *Singular I* (D. Mass. June 10, 2020), ECF No. 49 (emphasis added).

Apart from confirming that its new claims abandon the prior claims' purported inventive feature, Singular's new complaint fails to meet the *Iqbal*/*Twombly* pleading standard as elucidated by the Federal Circuit just last year. Under the decision in *Bot M8 LLC v. Sony Corp. of America*, Singular "need not prove [its] case at the pleading stage," but its claims are subject to dismissal if they "plead[] facts that are inconsistent with the requirements of its claims." 4 F.4th 1342, 1346 (Fed. Cir. 2021). Singular's complaint, even when read in the light most favorable to Singular, suffers such inconsistency, and dismissal is a necessary consequence.[3]

The claims that Singular now asserts are drawn to a floating-point number format plus some additional limitations. But Singular's pleading as to at least two of those additional limitations—"local" memory and "input-output unit"—fails to meet the pleading standard set by the Supreme Court and the Federal Circuit because Singular (i) does not directly allege that the accused products meet those claim limitations; and (ii) actually pleads facts that are inconsistent with the requirements of the claims.

"Local" Memory. The complaint does not even directly address the "local" memory limitation of the claims. Furthermore, a Google webpage that Singular cites on allegations of memory generally—but not as to "local" memory specifically—undercuts rather than supports Singular's allegation. The document illustrates the use of memory that is *external* to all the processing elements that is not even connected to any of them, and thus, decidedly not "local" to any processing element. If Singular sought to contend that the depicted memory meets the limitation of being "local to" a processing element, that would conflict with Singular's representation to this Court that "the specification states that a 'processing element' *comprises* an

[3] Concurrent with this Motion, Google has filed a motion to stay (ECF No. 12) that, if granted, would defer the need to take any action on the instant Motion at least until after conclusion of the stay.

arithmetic circuit paired with a memory circuit."[4] Although made in connection with the patents Singular asserted previously, the statement applies with equal force here because these patents have the same specification as the patents at issue in the prior case.

Input-Output Unit. Singular's allegations are similarly deficient as to the "input-output unit" limitation in the claims. The asserted claims all require an "input-output unit" that is (1) part of the claimed computing chip and (2) connected via an interface to both (a) a host computer and (b) at least one processing element. Singular makes no allegations that even purport to identify a structure in the TPU that meets these "input-output unit" claim limitations. And again, Google material that Singular cites undercuts rather than supports its allegations. Singular cites a TPU diagram showing that the unit containing the accused processing elements, *i.e.*, the MXU, is *not* connected to the host computer. Further, the image Singular cites in association with the only mentions of the words "input-output" in the complaint does not identify "input-output" on an accused computing chip; it thus could not satisfy the "input-output unit" limitation.

II. BACKGROUND

Singular alleges that versions 2, 3, and 4 of Google's tensor processing units ("TPUs") infringe claim 10 of the '616 Patent (which is dependent on claims 7 and 8) and claim 1 of the '775 Patent. ECF No. 1 ("Compl.") ¶¶ 25 n.1, 39-116. Both patents recite alleged inventions for "Processing with Compact Arithmetic Processing Element." *Id.* ¶¶ 20, 21. Because these patents are continuations of the three patents that Singular previously asserted against Google, they all share the same specification. Consistent with Singular's prior assertions about patent novelty, the "Summary" in the common specification refers to embodiments with "'low precision' processing elements [that] perform arithmetic operations which produce results that frequently differ from

[4] Pl.'s Preliminary Claim Construction Brief at 6, *Singular I*, ECF No. 112.

exact results by at least 0.1% (one tenth of one percent)." *See, e.g.*, Ex. A ("'616 Patent") at 2:16-19.[5]

The claims in the newly asserted patents abandon that purported point of novelty, however, and include no limitation directed to error. Instead, the new claims are drawn to a processing element that uses a floating-point number format with a maximum mantissa size and a minimum exponent size,[6] along with other physical limitations directed to the processing elements, including requirements of "local" memory plus an input-output unit that provides connectivity between the host computer and the processing element(s). *See, e.g.*, '616 Patent, claims 7 & 8; Ex. B ("'775 Patent"), claim 1.

A. The asserted claims require processing elements that each comprise a memory that is "local" to each processing element.

Both claims asserted in the complaint—claim 10 of the '616 Patent and claim 1 of the '775 Patent—are directed to a "computing system" that comprises a "host computer" and "a computing chip"; the chip in turn comprises certain additional elements, including certain processing elements. Compl. ¶¶ 25, 26. Each "first processing element" in '616 Patent claim 10, and all the processing elements in claim 1 of the '775 Patent, require a "memory" that is "local" to the processing element. The claim language sets forth these requirements as follows:

[5] The Court may consider the patent specifications in deciding the instant Motion because they are proper subjects of judicial notice. *See Secured Mail Sols., LLC v. Universal Wilde, Inc.*, 873 F.3d 905, 913 (Fed. Cir. 2017) (noting that "the claims and the patent specification" are "matters properly subject to judicial notice or by exhibit"). The patents share a common specification; citations herein are to the '616 patent.

[6] The Federal Circuit has explained the terms "mantissa" and "exponent" in the context of floating-point format as follows: "[D]ata is represented by the product of a fraction, or mantissa, and a number raised to an exponent. For example, a number n can be represented in base 10 by $n = m \times 10^e$, where m is the mantissa and e is the exponent. If m equals 2 and e equals 1, n equals 20; if m equals 2 and e equals -1, then n equals 0.2." *Silicon Graphics, Inc. v. ATI Techs., Inc.*, 607 F.3d 784, 786 (Fed. Cir. 2010).

- Claim 10 of the '616 Patent requires: "a plurality of memory units, wherein each of the plurality of first processing elements is associated with a corresponding one of the plurality of memory units, and wherein each of the plurality of memory units is *local* to its associated one of the plurality of first processing elements." Compl. ¶ 25 (emphasis added).[7]

- Claim 1 of the '775 Patent requires: "a first memory *local to* the first edge processing element; a second memory *local to* the second edge processing element; a third memory *local to* the first interior processing element; a fourth memory *local to* the second interior processing element." *Id.* ¶ 26 (emphasis added).

In the prior case, Singular addressed what the identical specification requires in a "processing element": "a 'processing element' *comprises* an arithmetic circuit paired with a memory circuit." Pl.'s Preliminary Claim Construction Brief at 6, *Singular I*, ECF No. 112 (emphasis added). Although the claim term "LPHDR execution unit" was at issue there, both parties had agreed that the claimed LPHDR execution units were "processing element[s]." *See* Pl.'s Reply Claim Construction Brief at 11-13, *Singular I*, ECF No. 135. The parties disputed whether all embodiments of the claimed LPHDR execution unit need to have a paired memory. Consistent with its position quoted above, Singular said they did.

B. The asserted claims require an "input-output unit" that is connected both to the processing elements and to the host computer.

The asserted claims require an "input-output unit" that is connected to the processing elements and a host computer. Compl. ¶¶ 25 ('616 Patent, claim 10), 26 ('775 Patent, claim 1). In pertinent part, claim 10 of the '616 patent requires "an input-output unit connected to each of the first subset of the plurality of first processing elements[,]" and "a host connection at least partially connecting the input-output unit with the host computer." *Id.* ¶ 25. Claim 1 of the '775 Patent

[7] As noted in Singular's complaint: "Claim 10 of the '616 patent is a dependent claim; it depends from claim 8, which in turn depends from independent claim 7." Compl. ¶ 25 n.1. Thus, the complaint quotes claim 10 "in independent form, to include the limitations of claims 7 and 8 from which it depends[,]" and Google does the same here. *Id.*

similarly requires, in pertinent part, "an input-output unit connected to the first edge processing element and the second edge processing element[,]" and "a host connection at least partially connecting the input-output unit with the host computer." *Id.* ¶ 26. Furthermore, the claims require that the input-output unit be part of the "computing chip" rather than part of the host computer or some other portion of the claimed computing system. *See id.* ¶¶ 25, 26.

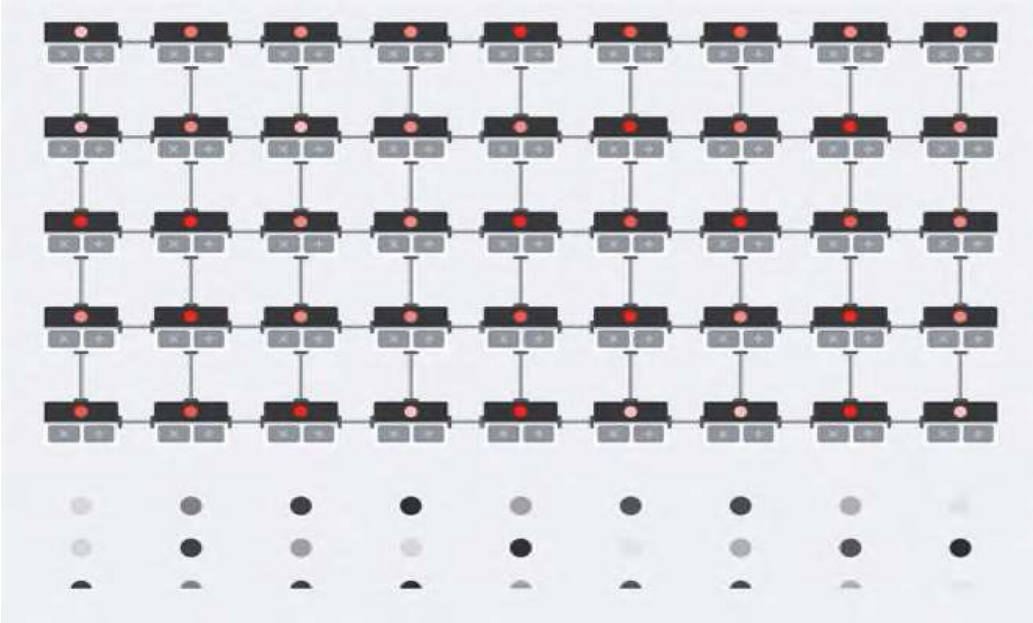
***

In sum, the asserted claims require architecture that includes, as part of the claimed computing chip: (i) individual processing elements, (ii) memory units local to each processing element, and (iii) an input-output unit connected to both specific processing elements and a host computer.

C. Singular identifies individual units that perform multiplication within the matrix multiplication unit as the accused processing elements.

The complaint alleges that Google's TPUs (versions 2, 3 & 4) infringe each limitation of the asserted claims. Compl. ¶¶ 25, 26. The complaint describes a TPU as including "a plurality of cores, each of which includes a Matri[x] Multiply Unit ('MXU') that runs matrix multiplications, a Vector Processing Unit ('VPU') and a Scalar Unit." *Id.* ¶ 48.

The complaint further alleges that the accused processing elements are within the MXUs. It states that "[e]ach TPU core includes at least one MXU that includes horizontally and vertically interconnected processing elements arranged in systolic arrays[,]" and that "each [MXU] processing element comprises an arithmetic unit that performs multiplication." *Id.* ¶¶ 51, 53. The complaint also alleges that each MXU has "16,384 processing elements." *Id.* ¶ 112. The complaint includes the following illustration to reflect the accused "processing elements arranged in systolic arrays":



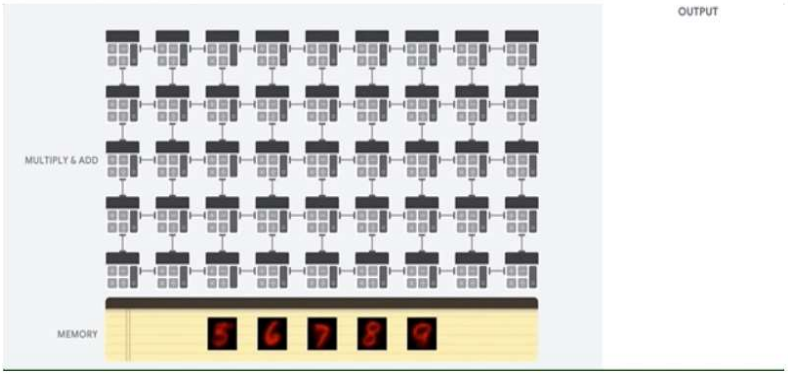
Compl. ¶ 51.

D. Singular makes general allegations about memory without specifically identifying its location, and the only Google document Singular cites identifies memory *outside* the accused processing element array.

As to the "local" memory limitations, noticeably absent from the complaint is any direct or specific allegation regarding a memory local to any processing element. Rather, the complaint states more generally that "[e]ach of the MXU processing elements has an associated memory." *See* Compl. ¶¶ 54, 109. The complaint then describes the alleged "associated memory" as being "used, for example to store 'weights' or 'parameters' as part of algorithms that relate to neural networks." *Id.* But the complaint does not allege what or where that memory is, whether it is part of the accused processing elements, or how it is connected to the processing element or paired with its arithmetic unit. The complaint does cite a Google webpage and then provides a parenthetical quote regarding the memory's functioning (but not its structure): "https://cloud.google.com/tpu/docs/beginners-guide ('the TPU loads the parameters from memory

into the matrix of multipliers and adders')." *See id.* Exhibit C is a screenshot from the archived webpage cited in complaint:[8]

Let's see how a systolic array executes the neural network calculations. At first, the TPU loads the parameters from memory into the matrix of multipliers and adders.



Then, the TPU loads data from memory. As each multiplication is executed, the result will be passed to the next multipliers while taking the summation at the same time. So the output will be the summation of all multiplication results between data and parameters. During the whole process of massive calculations and data passing, no memory access is required at all.

Ex. C. The webpage shows memory (highlighted in yellow) that is separate from the MXU processing elements directly above it. Likewise, other figures cited in the complaint also show only memory outside the MXU. *See* Compl. ¶¶ 45, 48.

[8] Exhibit C is a screenshot of the portion of the archived webpage cited in the complaint, retrieved from The Wayback Machine at https://web.archive.org/web/20211009235618/https://cloud.google.com/tpu/docs/beginners-guide. The Court may consider the archived webpage because it is incorporated by reference in the complaint. *See, e.g.*, *Watterson v. Page*, 987 F.2d 1, 3-4 (1st Cir. 1993) (noting that materials "expressly incorporated" in the complaint, or "central" to a plaintiff's claim, or "sufficiently referred to in the complaint" may be considered in deciding a motion to dismiss). Here, the complaint expressly incorporates the webpage by reference by citing it as evidence of infringement and directly quoting from it. *See* Compl. ¶¶ 54, 109; *see also Clorox Co. Puerto Rico v. Proctor & Gamble Com. Co.*, 228 F.3d 24, 32 (1st Cir. 2000) (noting that courts "'may properly consider the relevant entirety of a document integral to or explicitly relied upon in the complaint'" in considering a Rule 12(b)(6) motion to dismiss) (citation omitted).

E. Singular does not identify any specific structure in the accused TPU as the "input-output unit."

The complaint references an "input-output Host VM CPU" or "host input-output CPU" connected to a TPU; it does not reference an "input-output unit," other than quoting the asserted claims. *See* Compl. ¶¶ 46, 103. Critically, the complaint does not allege that the accused products contain an input-output unit that is connected to at least one accused processing element (which, as discussed above, are the units inside the MXU performing arithmetic). Nor, as a further matter, does the complaint allege that any such input-output unit is part of the accused computing chip or that it is also partially connected to the host computer. In the only two places where it uses the phrase "input-output," the complaint includes an image allegedly showing the "input-output Host VM CPU" or "host input-output CPU" connected to the TPU. *See* Compl. ¶¶ 46, 103; *see also infra* Part.IV.B.2 at 19. But the complaint does not have any allegations tying that image to any specific infringement allegations or claim limitations.

Moreover, the complaint also includes another block diagram image of the TPUv2, with some related text. That image shows that the MXU (illustrated in blue) is connected only to the VPU and the interconnect router, neither of which Singular has suggested is either the host computer or an input-output unit. *See* Compl. ¶ 64; *see also* Appendix (reproducing complaint image); *infra* Part.IV.B.2 at 18.

III. LEGAL STANDARD

Dismissal pursuant to Federal Rule of Civil Procedure 12(b)(6) is required where a complaint does not "contain sufficient factual matter, accepted as true, to 'state a claim to relief that is plausible on its face.'" *Ashcroft v. Iqbal*, 556 U.S. 662, 678 (2009); *see also Bell Atl. Corp. v. Twombly*, 550 U.S. 544, 555 (2007) ("Factual allegations must be enough to raise a right to relief above the speculative level."). Allegations "that are 'merely consistent with' a defendant's

liability" do not establish facial plausibility. *Iqbal*, 556 U.S. at 678 (quoting *Twombly*, 550 U.S. at 557). Further, factual allegations in patent cases that "are actually *inconsistent* with and contradict infringement[ ] . . . are likewise insufficient to state a plausible claim." *Bot M8*, 4 F.4th at 1354. "[M]ere recitation of claim elements and corresponding conclusions, without supporting factual allegations, is insufficient to satisfy the *Iqbal/Twombly* standard." *Id.* at 1355.

In sum, if the complaint fails to assert "'factual allegations, either direct or inferential, respecting each material element necessary to sustain recovery under some actionable legal theory,'" dismissal under Rule 12(b)(6) is appropriate. *Gagliardi v. Sullivan*, 513 F.3d 301, 305 (1st Cir. 2008) (citation omitted).

IV. ARGUMENT

Singular fails to plead a plausible infringement claim, even accepting its non-conclusory factual allegations as true, for several independent reasons: (1) the complaint does not directly address—conclusorily or otherwise—the "local" memory limitations, and the material the complaint cites as to memory contradicts any allegation that the memory identified in the accused device is local; and (2) the complaint does not directly address the "input-output unit" limitation, and the material the complaint cites shows that the accused processing elements do not connect to any host computer via an input-output unit, as the claims require. Dismissal is independently warranted on either ground.

A. Singular fails to address the "local" memory limitation that is in all asserted claims, and the evidence it cites as to the alleged memory component of the accused TPU is inconsistent with and contradictory to infringement.

The asserted claims require that each processing element has a "local" memory. But the complaint contains *no* allegations regarding the "local" memory limitation. Further, the evidence Singular cites regarding what it alleges to be the memory component of the accused device demonstrates non-infringement as to the "local" memory limitation.

1. Singular has failed to plead infringement plausibly, because it does not make any allegation regarding the "local" memory limitation, which appears in both claims asserted in the complaint.

The asserted claims all require memory "local" to each processing element. *See* Compl. ¶¶ 25 (quoting claim 10 of the '616 Patent), 26 (quoting claim 1 of the '775 Patent). But the complaint does not even attempt to identify any structure in the TPUs that constitutes memory "local" to a processing element. In fact, the complaint contains *no* direct allegations—conclusory or factual—asserting that the accused devices include an array of processing elements and corresponding memory units that are *local* to each individual element. Rather, the *only* allegation even arguably directed to "local" memory is a "catch-all" allegation that simply recites the claim language and states that the accused *TPUs* meet all the limitations in the asserted claims. *See* Compl. ¶¶ 25, 26. Notably, this allegation does not even identify the accused processing elements, much less what "local" memory those processing elements have.

Even if credited as an allegation directed to local memory, this allegation at best merely parrots the claim language without doing more, and as such is insufficient to state a plausible infringement claim. Indeed, the Federal Circuit has made clear that "mere recitation of claim elements and corresponding conclusions, without supporting factual allegations, is insufficient to satisfy the *Iqbal/Twombly* standard." *See Bot M8*, 4 F.4th at 1355; *see also Swirlate IP LLC v. Keep Truckin, Inc.*, No. 20-1283-CFC, 2021 WL 3187571, at *2 (D. Del. July 28, 2021) (dismissing complaint for failing to satisfy *Iqbal*/*Twombly* where plaintiff failed to sufficiently allege how the accused product infringed the asserted limitations); *People.ai, Inc. v. SetSail Techs., Inc.*, No. C 20-09148 WHA, 2021 WL 2333880, at *2-5 (N.D. Cal. June 8, 2021) (dismissing claims for direct infringement because complaint failed to plead facts explaining how the accused product infringed the asserted limitations); *NovaPlast Corp. v. Inplant, LLC*, No. 20-7396 (KM) (JBC), 2021 WL 389386, at *7-8 (D.N.J. Feb. 3, 2021) (same).

Singular's only allegation relating to both memory and the accused processing element is a conclusory allegation about *associated* memory, where it states that "[e]ach of the MXU processing elements has an associated memory." Compl. ¶¶ 54, 109. That allegation, even if proven true, would not establish that the alleged "associated memory" component is "local" to each accused processing element, as required by all of the asserted claims. *See* Compl. ¶¶ 25, 26. Singular cannot state a plausible claim of infringement by simply pointing to memory that is found *anywhere* on the accused device. Instead, for the "local" limitation to have any meaning, Singular's complaint must identify memory that is actually "local" to each processing element. Otherwise, Singular is alleging that any memory located on a device somewhere in the computing system meets this claim element, which cannot be correct even drawing all inferences in favor of Singular. Nearly all modern computers have memory somewhere; the claims require more than that, and Singular must offer allegations that plausibly identify a memory *local* to each processing element. *See, e.g.*, *Swirlate IP*, 2021 WL 3187571, at *2; *People.ai, Inc.*, 2021 WL 2333880, at *2-5; *NovaPlast Corp.*, 2021 WL 389386, at *7-8.[9]

As to the limitation of "local" memory, Singular's reference to "associated memory" does not meet even the standard it set when making representations to the Court in the prior case about what a processing element comprises under the specification. There, Singular said that a

[9] Moreover, while memory being associated with a processing element is a separate claim element from it being local (*see, e.g.*, '616 Patent claim 1), Singular's entirely conclusory allegation would be insufficient even as to that separate element. *See Bot M8*, 4 F.4th at 1355 (mere recitation of claim elements and corresponding conclusions, without supporting factual allegations, is insufficient to satisfy the *Iqbal/Twombly* standard). The complaint includes no facts in support of the allegation regarding "associated" memory. Instead, the complaint describes the alleged "memory" as being "used, for example to store 'weights' or 'parameters' as part of algorithms that relate to neural networks." Compl. ¶¶ 54, 109 (citing https://cloud.google.com/tpu/docs/beginners-guide (*see* Ex. C)). That allegation merely provides an example of the alleged memory's functionality, and does not explain how the alleged memory is associated with any individual processing element.

processing element within the meaning of the patent specification must *comprise* a memory. Pl.'s Reply Claim Construction Brief at 12, *Singular I*, ECF No. 135. Singular has made no allegation identifying any memory that would meet those self-imposed requirements.

In sum, because Singular fails to plead facts that, even if proven, would establish that the accused devices satisfy the "local" memory limitations, dismissal is warranted. *See Bot M8*, 4 F.4th at 1353 ("There must be some factual allegations that, when taken as true, articulate why it is plausible that the accused product infringes the patent claim.").

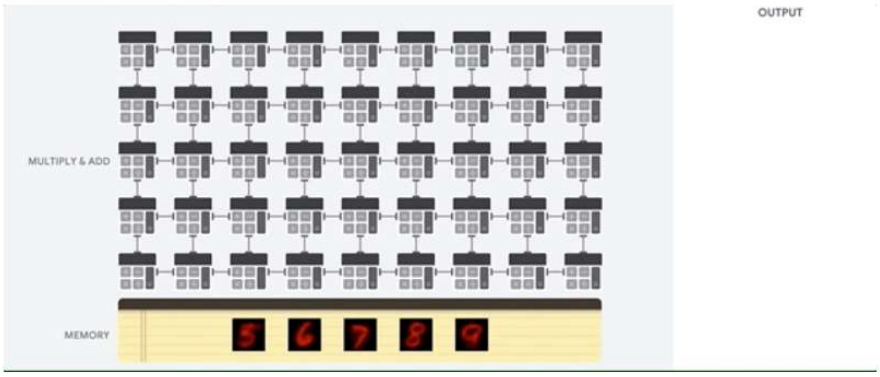
2. The only document that Singular cites in its complaint regarding memory is inconsistent with infringement as to the "local" memory limitations.

Even if the one document that Singular cites when discussing "associated memory" was somehow credited as sufficient to constitute an allegation directed to the separate "local" memory limitations, that document actually undercuts rather than supports an infringement claim, because the only memory that the document describes could not constitute memory "local to" a processing element under any interpretation of that term. Rather, the cited document shows that the memory described therein is not "local to" any accused processing element.

As previously discussed, Singular accuses the multiplication units within the MXU as the claimed "processing elements," and the asserted claims require that each such "processing element" must also have a memory local to it. While making no direct mention of local memory, Singular alleges that "[e]ach of the MXU processing elements has an associated memory." *See* Compl. ¶¶ 54, 109. In an apparent attempt to provide support for that allegation, the complaint cites a Google webpage. *See* Compl. ¶¶ 54, 109 (citing "https://cloud.google.com/tpu/docs/beginners-guide ('the TPU loads the parameters from memory into the matrix of multipliers and adders')."); *see also* Ex. C (archived webpage cited in complaint). The cited page shows an array of processing elements with the label "Multiply & Add"

to the left; these are what Singular accuses as the claimed processing elements. *See* Compl. ¶¶ 51, 53. But the *memory* shown on that webpage is separate from not only each individual processing element but also the array of processing elements as a whole. *Id.*; *see also* Compl., ¶¶ 45, 48 (showing only memory entirely separate from the MXU). Thus, this memory could not constitute a memory local to a processing element under any interpretation supported by Singular's own representation that the processing element must have its own memory.

Let's see how a systolic array executes the neural network calculations. At first, the TPU loads the parameters from memory into the matrix of multipliers and adders.



Then, the TPU loads data from memory. As each multiplication is executed, the result will be passed to the next multipliers while taking the summation at the same time. So the output will be the summation of all multiplication results between data and parameters. During the whole process of massive calculations and data passing, no memory access is required at all.

Ex. C.[10] With the "local" memory allegation disproven by the very evidence Singular cites, Singular's complaint is devoid of *any* allegation that even arguably could render it plausible that the TPU has the claimed memory local to a processing element. *Bot M8*, 4 F.4th at 1355 (mere

[10] The archived webpage now redirects to https://cloud.google.com/tpu/docs/intro-to-tpu, which includes the same graphic. The current page identifies the memory component, illustrated in yellow, as high bandwidth memory ("HBM"), noting: "To perform the matrix operations, the TPU loads the parameters from HBM memory into the MXU." Singular's complaint references "HBM," but not in the context of the local memory limitation. *See* Compl. ¶ 49 ("Each TPU chip on the board is external to the other three chips and includes at least one host interface to a host computer and [HBM]."); *see also id.* ¶¶ 45, 48 (graphics illustrating HBM that is external to the MXU).

recitation of claim elements and corresponding conclusions, without supporting factual allegations, is insufficient to satisfy the *Iqbal/Twombly* standard).

Moreover, the narrative text accompanying the illustration in the webpage supports the lack of any direct connection between the memory and the accused processing elements. It says that "the TPU loads the parameters from memory into the *matrix* of multipliers and adders," but does not say that there is any direct connection between the memory and any individual multiplier or adder. *See* Ex. C. The graphic shows a common memory (illustrated in yellow), which as the accompanying text and visuals show, the individual processing elements do not directly access. Rather, as the text notes, "[n]o memory access is required during the matrix multiplication process." *Id.* The matrix multiplication process is what the accused processing elements are used for. So, the cited document explains that the processing elements do not need to access *any* memory when performing the function they were designed for.

Singular's inconsistent allegations of infringement are similar to the allegations deemed deficient in *Bot M8*. There, the plaintiff alleged that the accused device's authentication program was located on the motherboard, which was inconsistent with the requirement of the asserted claim limitation that the authentication program and the game program be stored together, *separate* from the motherboard. *Bot M8*, 4 F.4th at 1353-54. The Federal Circuit therefore affirmed the district court's dismissal, noting that the plaintiff's allegations rendered its "infringement claim not even possible, much less plausible." *Id.* at 1354.

Bot M8 compels dismissing Singular's claims of infringement. Assuming the truth of the "memory" allegations and drawing all inferences in Singular's favor, the evidence Singular cites demonstrates that the memory is not "local" to each MXU processing element, but is instead a

separate memory not even connected to the individual processing elements and outside the MXU itself. *See* Compl. ¶¶ 54, 109; Ex. C. Accordingly, dismissal is warranted.

B. Singular fails to address the "input-output unit" limitation that is in all asserted claims, and the Google materials it references in the complaint are inconsistent with and contradictory to infringement.

As with the "local" memory limitations, the complaint fails to plead a plausible claim of infringement as to the required "input-output unit" limitation for two independent reasons. First, the complaint makes no allegation that the accused TPUs include an "input-output unit" that is connected to both a host computer and to an accused processing element, as the claims require. Second, the Google materials that Singular does cite do not support the allegations of infringement, because they show that there is no possible connection between the accused processing elements, which sit inside the MXU, and any host computer.

Claim 10 of the '616 Patent and claim 1 of the '775 Patent both require "a host connection at least partially connecting the input-output unit with the host computer." Compl. ¶¶ 25, 26. The claimed "input-output unit" must also be "connected" to specific processing elements; specifically, claim 10 of the '616 patent requires "an input-output unit connected to each of the first subset of the plurality of first processing elements[,]" and claim 1 of the '775 Patent similarly requires "an input-output unit connected to the first edge processing element and the second edge processing element." *Id.* Moreover, the claimed input-output unit is a sub-part of the claimed "*computing chip*" sub-part, rather than the "host computer"[11] sub-part:

> A computing system, comprising:
> a host computer;

[11] Fig. 1 of the specifications illustrate an embodiment of the invention that includes a separate "Host" and "Input/Output Unit." *See, e.g.*, '616 Patent Fig. 1. The specifications describe the "Host" as being "responsible for overall control of the computing system," and "perform[ing] the serial, or mostly serial, computation typical of a traditional uni-processor." *See, e.g.*, '616 Patent, 8:29-31.

> *a computing chip comprising:*
> . . .
> an input-output unit connected to each of the first subset of the plurality of first processing elements;
> . . .
> a host connection at least partially connecting the input-output unit with the host computer; . . . .

Compl. ¶ 25 (quoting '616 Patent); *see also id.* ¶ 26 (quoting nearly identical language from '775 Patent). But, although the complaint identifies the TPU as the alleged "computing chip," it fails to identify any "input-output unit" of that TPU as allegedly satisfying the claim limitations.

1. Singular has failed to plead infringement plausibly because it does not make any allegation that even attempts to identify an "input-output unit" meeting the claim limitations.

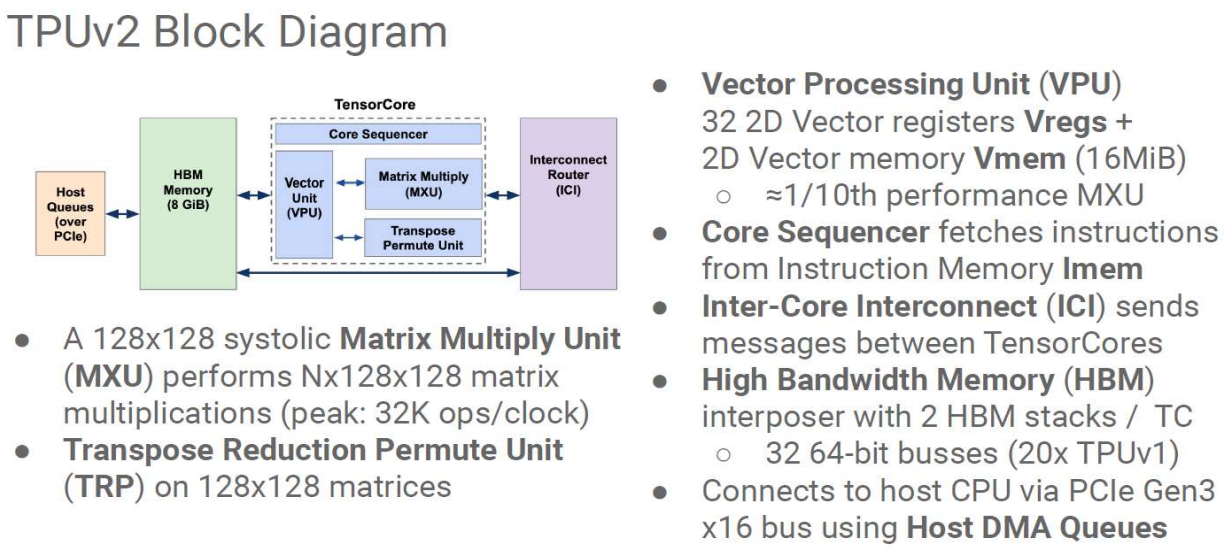
Here, the complaint does not identify an input-output unit in the TPU at all, much less an input-output unit connected both a host computer and to the accused processing elements, *i.e.*, the units within the MXU that perform arithmetic. The closest the complaint comes to making such an allegation is when it alleges that "[e]ach TPU chip on the board is external to the other three chips and includes at least one host interface to a host computer and High Bandwidth Memory ('HBM')." *See* Compl. ¶ 49. But Singular does not allege that "host interface" meets the "input-output unit" limitation.

Moreover, the input-output unit must connect to an accused processing element, and the complaint does not even have the slightest suggestion that the host interface connects to a processing element. Singular cannot state a plausible claim by merely referencing a feature of an accused device that contains similar wording to an asserted limitation without setting forth facts that show *how* or *why* that feature infringes. *See, e.g.*, *Bot M8*, 4 F.4th at 1353 (noting that a plaintiff must "articulate why it is plausible that the accused product infringes the patent claim"); *People.ai*, 2021 WL 2333880, at *2-5 (dismissing claims for failing to allege "how or why" the accused product infringed the asserted claims). Indeed, simply identifying a graphic that uses the

word "input-output" cannot plausibly allege infringement of this element, which adds specific limitations to the input-output unit, because input and output, like memory, exist in almost every modern computer. *See, e.g.*, '616 Patent, 4:21-28 (identifying input and output in prior art field programmable array devices).

2. The Google materials Singular cites in the complaint are inconsistent with infringement as to the "input-output unit" limitations.

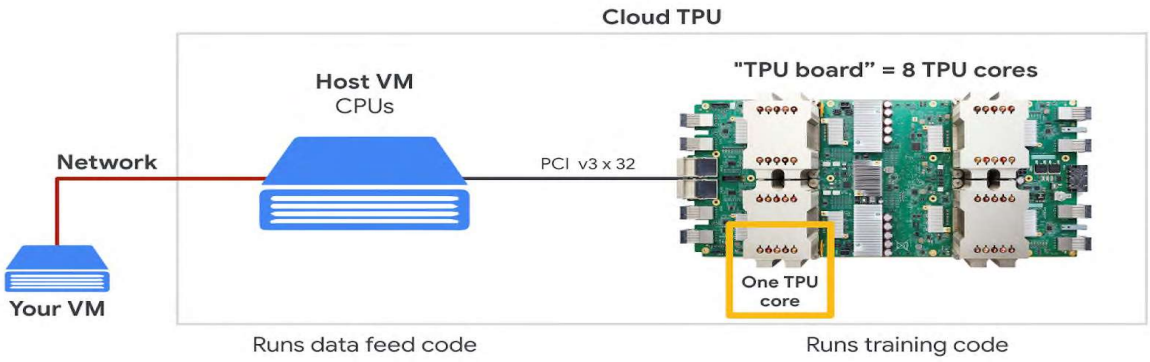
Similar to Singular's allegations regarding the asserted "memory" limitations, the complaint demonstrates only non-infringement as to the asserted "input-output unit" limitation. First, the complaint cites a "TPUv2 Block Diagram" that allegedly shows how, "[a]ccording to Google, TPUs perform multiplication":



Compl. ¶ 64; *see also* App. The TPU v2 Block Diagram shows there is no connection between the MXU, which contains the accused processing elements, and the host CPU. Rather, the diagram shows that the MXU connects only to the Vector Unit (VPU) and the Interconnect Router, neither of which is alleged to be the claimed input-output unit or alleged to provide a connection to an alleged host computer. Given that the accused processing elements are further embedded *within*

the MXU, this diagram shows that those processing elements, even if they were connected to an unidentified and unaccused input-output unit, could not be shown to have a connection to a host computer, because the only unit that this diagram shows the MXU connected to are the VPU and an interconnect router. Thus, dismissal is warranted because Singular's allegations render its "infringement claim not even possible, much less plausible." *See Bot M8*, 4 F.4th at 1354.

Second, the one place in the complaint that actually uses the phrase "input-output" also contradicts Singular's infringement claims, if Singular were to rely on that allegation as the basis for having sufficiently plead the input-output unit element. Singular alleges: "Google's Cloud Platform ('GCP') comprises a computer system having at least one input-output Host VM CPU connected to at least one TPU board having a plurality of TPU cores"; and "[t]he accused TPUs comprise circuit boards that are connected to a host input-output CPU." Compl. ¶¶ 46, 103. Singular also includes the following graphic from a Google document:



Id. But like the narrative allegations, the image does not show an input-output unit that is (1) a sub-part of the accused computing chip (*i.e.*, the TPU); (2) connected to the processing elements (*i.e.*, located in the MXU); and (3) separate from and connected to the host computer—all of which is required by the asserted claims. The image instead shows a "Host VM CPU[]" that is: (1) *external* to the TPU chip (*i.e.*, not a sub-part of the same chip); (2) *not* connected to the MXU processing elements; and (3) *part of* the host computer.

Accordingly, dismissal is warranted because the structure of the accused TPU, as alleged, is inconsistent with Singular's infringement claims. *See Bot M8*, 4 F.4th at 1354.

V. CONCLUSION

For the foregoing reasons, the complaint should be dismissed for failure to state a claim.

Respectfully submitted,

Date: February 11, 2022

/s/ Nathan R. Speed

Gregory F. Corbett (BBO #646394)
gcorbett@wolfgreenfield.com
Nathan R. Speed (BBO # 670249)
nspeed@wolfgreenfield.com
Anant K. Saraswat (BBO #676048)
asaraswat@wolfgreenfield.com
Elizabeth A. DiMarco (BBO #681921)
edimarco@wolfgreenfield.com
WOLF, GREENFIELD & SACKS, P.C.
600 Atlantic Avenue
Boston, MA 02210
Telephone: (617) 646-8000
Fax: (617) 646-8646

Robert Van Nest*
rvannest@keker.com
Michelle Ybarra*
mybarra@keker.com
Eugene M. Paige*
epaige@keker.com
Andrew Bruns*
abruns@keker.com
Vishesh Narayen*
vnarayen@keker.com
Anna Porto*
aporto@keker.com
Deeva Shah*
dshah@keker.com
Stephanie J. Goldberg*
sgoldberg@keker.com
KEKER, VAN NEST & PETERS LLP
633 Battery Street
San Francisco, CA 94111-1809
(415) 391-5400

Michael S. Kwun*
mkwun@kblfirm.com
Asim Bhansali*
abhansali@kblfirm.com
KWUN BHANSALI LAZARUS LLP
555 Montgomery Street, Suite 750
San Francisco, CA 94111
(415) 630-2350

Matthias A. Kamber*
matthiaskamber@paulhastings.com
PAUL HASTINGS, LLP
101 California Street
Forty-Eighth Floor
San Francisco, CA 94111

Counsel for Defendant Google LLC
**motions for pro hac vice to be filed*

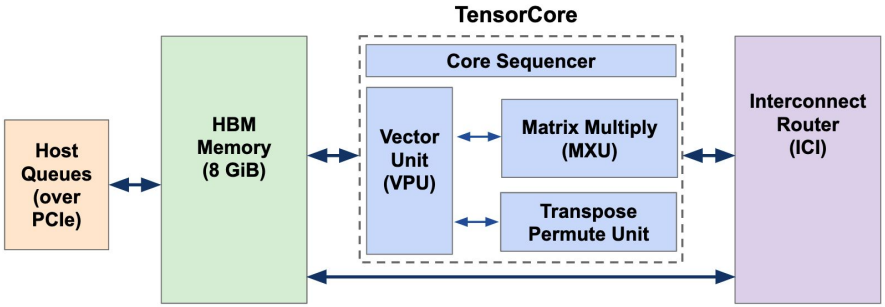
CERTIFICATE OF SERVICE

I certify that this document is being filed through the Court's electronic filing system, which serves counsel for other parties who are registered participants as identified on the Notice of Electronic Filing (NEF). Any counsel for other parties who are not registered participants are being served by first class mail on the date of electronic filing.

/s/ Nathan R. Speed
Nathan R. Speed

Appendix

TPUv2 Block Diagram



- A 128x128 systolic **Matrix Multiply Unit** (**MXU**) performs Nx128x128 matrix multiplications (peak: 32K ops/clock)
- **Transpose Reduction Permute Unit** (**TRP**) on 128x128 matrices

- **Vector Processing Unit** (**VPU**)
  32 2D Vector registers **Vregs** +
  2D Vector memory **Vmem** (16MiB)
  - ≈1/10th performance MXU
- **Core Sequencer** fetches instructions from Instruction Memory **Imem**
- **Inter-Core Interconnect** (**ICI**) sends messages between TensorCores
- **High Bandwidth Memory** (**HBM**) interposer with 2 HBM stacks / TC
  - 32 64-bit busses (20x TPUv1)
- Connects to host CPU via PCIe Gen3 x16 bus using **Host DMA Queues**